CS 327E Milestone 4 due Sunday, 02/23.

Perform these modeling tasks to improve the quality and usability of the data in your `dataset1`.

1. Create a new Jupyter notebook named `<source>_modeled.ipynb` where `<source>` is the source of your `dataset1`. Implement the following logic in your notebook.

2. Create a new BQ dataset to store your modeled tables for `dataset1`. The dataset name
   should follow the naming convention `<source>_modeled` where `<source>` is the source of the data. For example, fda_modeled.

3. Create new tables in your modeled dataset by applying the design principles learned in class:
   - split any staging tables that contain more than one entity into separate tables.
   - join staging tables that store different attributes belonging to the same entity.
   - union staging tables that store distinct records belonging to the same entity type.
   - identify a primary key (PK) for each modeled table.
   - check for the presence of duplicate records in each modeled table. If duplicate records exist, give an example of the duplicate records in question in a file named `TRANSFORMS.txt.`

4. Identify relationships between the modeled table:
   - connect the tables in the diagram using the appropriate relationship type.
   - check for any referential integrity violations. If violations exist, give an example of the violations in question in the file named `TRANSFORMS.txt.`

5. For each field in the modeled tables, choose a primitive data type that most precisely represents its domain of values:
   - if the field is of type `STRING` and it stores `INTEGER`, `NUMERIC`, `DATE` or `TIMESTAMP` values, cast its type to the most fitting type.
   - if the field is of type `INTEGER` and it stores a `DATE` or `TIMESTAMP` value, cast its type to the most fitting type.
   - if the field is of type `TIMESTAMP` and the values it stores are of type `DATE` (i.e. the time component is not being used), cast its type to `DATE`.
   - use BQ's **`CAST`** function to convert from one type to another.
   - if the **`CAST`** function returns an error, make a note of the field and the error in the file named `TRANSFORMS.txt`.

6. Create an ERD of the modeled tables in your `dataset1`:
   - The diagram should represent the current state of your modeled tables even if some design issues remain (e.g. duplicate records, foreign key violations)

- The diagram should include the set of field names and data types for each entity type and where applicable keys (PK, FK) for each entity type.
- The diagram should include valid relationships between entities.
- Name the ERD file `<source>_erd_modeled.pdf` where `<source>` is the source of your `dataset1`.

7. Rewrite your join queries from Milestone 3 to run over the modeled tables and update your notebook `<source>_joins.ipynb` with your code changes.

CS 327E Milestone 4 Rubric
Due Date: 02/23/20

For `dataset1`, identify all entity types in your tables, split additional entity types into their own tables, join tables belonging to the same entity type, and union all tables that share the same fields. All modeled tables should have an identified primary key unless otherwise justified in `TRANSFORMS.txt`. Values in the primary key should have no duplicates. String fields, if able to be casted to a more fitting type, should be. **-40** `<source>_modeled.ipynb` not found in repository **-10** `<source>_modeled` dataset not found in BQ project **-20** no primary keys identified in ERD and missing valid explanation in `TRANSFORMS.txt` **-10** marked primary keys contain duplicates **-10** each string field containing only `INTEGER, NUMERIC, DATE,` or `TIMESTAMP` not cast, up to **-40** partial credit is awarded for explanations in `TRANSFORMS.txt` **-10** each non-merged entity type, table with multiple entity types, or un-unioned tables containing the same data (i.e tables representing the same data across different years).	40
For `dataset1`, all child tables should have an identified foreign key unless orphaned rows contained in child tables and noted in `TRANSFORMS.txt`. **-30** no foreign keys identified on child tables in ERD **-20** foreign key relation is incorrect **-10** missing explanation in `TRANSFORMS.txt`	30
An ERD which contains detailed information for the fields in the modeled tables. Note that credit from other parts of the assignment may rely on this part. **-30** `./<source>_erd_modeled.pdf` not found in repository **-10** missing field types **-10** missing field names **-10** missing field keys **-5** incorrect keys marked	30
Update join queries to run on modeled tables. Make sure each statement runs properly. Save them into the same notebook, replacing the broken statements with their fixed counterparts. **-5** each erroneous SQL query, up to **-20**	

`submission.json` submitted into Canvas. Your project **will not** be graded without this submission. The file should have the following schema:
 	
`{`
`"commit-id": "your most recent commit ID from Github",`
`"project-id": "your project ID from GCP"`
	Required

`}`
 	
Example:
 	
`{`
`    "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",`
`    "project-id": "some-project-id"`
`}`	

Total Credit:	**100**